LA-UR-21-28460
Approved for public release; distribution is unlimited.

Title: Summer 2021 Internship Report

Author(s): Jayawardena, Handapangoda Mudalige Gavindya Nuwandi

Intended for: Report

Issued: 2021-08-24

Summer 2021 Internship Report

Gavindya Jayawardena
UIN: 01130618
hjaya001@odu.edu

CS669 Practicum
August 20, 2021

- I ensure that the information that Gavindya Jayawardena has discussed is appropriate for release and does not violate company regulations or confidences.
- I ensure that the report is an accurate description of the duties and functions of the assigned position during the work period.

Brian Cain	Brian Cain	8/24/2021
Name of the Supervisor	Signature	Date

I: Position Information

Title: Summer Research Intern at Los Alamos National Laboratory (LANL).

The internship was a 12 week program which started on 7th of June 2021. During this internship program, I worked as a Research Intern on a project which focused on feature extraction from scientific PDF documents using open-source, machine-learning-based tools.

Organization: Los Alamos National Laboratory, Los Alamos, NM

Los Alamos National Laboratory (LANL) is a United States Department of Energy national laboratory. It is located in Los Alamos, New Mexico, in the southwestern United States. Its mission is to solve national security challenges through scientific excellence. As a Federally Funded Research and Development Center, LANL aligns its strategic plan with priorities set by the Department of Energy's National Nuclear Security Administration. This year, LANL continued its student internship program for Summer 2021. Approximately 1500 students have joined LANL to work on various projects. Due to social distancing restrictions, most internships were limited to remote work off laboratory property.

Department: Institutional Scientific Content (ISC) Team, a sub division of Research Library

Supervisor: Brian Cain - Library Technology Professional

I worked under the supervision of Brian Cain throughout the internship program. Brian Cain is the leader of Institutional Scientific Content (ISC) team.

II: Duties/Responsibilities

Responsibility:

During the internship program, I was assigned to work with open-source, machine-learning-based tools such as GROBID for feature extraction on PDF documents. The goal is to extract metadata (title, authors, affiliations, abstracts, and keywords) from PDF documents in real-time

and process this data into a desired structure. This structured data can then be ingested into forms that could be used to populate library systems such as RASSTI (Review and Approval System for Scientific and Technical Information). Therefore, speed, accuracy, and completeness are key aspects of development. Once the extraction and data structuring work convincingly, the goal is to use the developed system for future integration in the production submission workflow for LANL's review and release system, RASSTI. This integration will help LANL researchers more easily submit their materials for review and also aid in the efficiency and accuracy of bibliographic descriptions of these documents for greater access by the LANL and DOE communities.

Responsibility Level: I was fully responsible for carrying out the project, which involved feature extraction on PDF documents using open-source, machine-learning-based tools, evaluating their speed and performance, and using selected tool to develop a prototype for metadata extraction and data structuring.

Major Activities: During the initial week of the internship program, I attended multiple training sessions such as Security Information Briefing and Working Safely during the COVID-19 Pandemic. Also, I received access to the institutional email address and the employee portal as well. Throughout this program, I attended meetings with my supervisor (Brian Cain) and the development team, development sessions, and meetings with the entire ISC team. The meetings with the development team were to update my progress and to obtain feedback to resolve issues or to improve the solution. The entire team meetings were held bi-weekly via Webex, where all team members updated their progress on their work. I contacted my supervisor via emails and Google chat throughout the internship. During the second week, I started working on my assigned project. To facilitate the development process of the project, my supervisor shared an

openly available corpus of PDF documents with me. To begin with, I first familiarized myself with GROBID by reading the documentation and exploring the functionalities of GROBID. I also learned about different functionalities of GROBID such as header extraction and parsing, parsing of names, parsing of affiliation, and full text extraction. Next, I focused on setting up the GROBID server locally. I found a Python client for GROBID REST services which is used to process PDFs.

In addition to GROBID, I discovered that PyPDF2 could be used extract metadata from PDFs as well. When I tested it out, unfortunately PyPDF2 extracts document information such as title and the author of the PDF, but not the complete list of authors or abstract or keywords, which our project need. Therefore, we did not utilize PyPDF2. I also discovered the SciWING, a scientific document processing toolkit which had ParsCit integrated. SciWING is built on PyTorch and it includes pre-trained models to extract features such as title, abstract, authors, affiliations and keywords from scientific documents. However, when I tested it out, it takes about one minute on average to extract features. Since speed is one of the key aspects of the development, we decided not to utilize SciWING as well.

After experimenting with PyPDF2 and SciWING, I proceeded with utilizing GROBID's Python client for feature extraction. From the GROBID web services, the two main web services I worked with are "processHeaderDocument" and "processFulltextDocument". These web services accepts a PDF file to extract features from the input PDF document, and convert it into a TEI XML format.

I first implemented a feature extractor based on GROBID's "processHeaderDocument" web service. I implemented the feature extractor such that it process the output and structure the features extracted, tabularly. The fields included in the table are file name, title, authors name

list, keywords list, abstract, and all affiliations list. I also curated a dataset with 20 sample PDFs and their features including title, abstract, authors, affiliations and keywords to perform quantitative evaluation. I created two different sets of ground truth, which is the ideal expected result in tabular format; one including affiliations of authors and another without affiliations of authors. In addition, I also created a JSON schema to store the extracted features including file name, title, abstract, keywords list, authors list with author name and affiliations. I compared ground truth with real-time GROBID extraction and calculated the accuracy of extracted titles (95%), authors (85%) , keywords (75%) , and abstracts (80%).

Upon evaluation, I noticed that sometimes the output of the GROBID gets saved with Unicode characters, yielding reduced accuracy. Upon discussing the initial implementation and the evaluation, the development team suggested to change the code to save the output without Unicode characters. In addition, we also decided not to proceed with the tabular structure. Instead to use JSON objects to store the output of GROBID. For the affiliations of authors, the development team suggested to use only the institution and not the raw affiliation which includes the street address as well. When I changed the JSON to have only the institution names as the affiliations, I discovered sometimes authors have affiliations and sometimes affiliations are not associated with the authors. I proposed to include a separate complete affiliations list regardless of the author in the JSON as well based on Brian's suggestion to allow users of RASSTI to select the affiliations from a drop drown in the interface.

Then I evaluated the JSON files. For this, I created a ground truth JSON files, which is the ideal expected result in JSON format from all 50 PDF files. The fields of the JSON objects are abstract, authors list [affiliations list, author first name, middle name, last name], file name, keywords list, title. Then I compared the JSON files of ground truth and JSON files of generated

output of GROBID. I calculated the accuracy of extracted titles (96%), keywords (86%), and abstracts (78%), authors first name (84%), authors middle name (78%), and authors last name(78%).

For the affiliation accuracy evaluations, I utilized affiliations only accuracy [48%], authors with affiliations accuracy [28%], micro-average precision [96.37%], macro-average precision [72.99%], micro-average recall [64.56%], and macro-average recall [62.59%]. Micro-average precision means that when a PDF is submitted, out of extracted affiliations of authors, 96.37% were actually correct. Micro-average precision is calculated by aggregating the contributions of PDFs and by not treating all PDFs equally. Macro-average precision means that when a PDF is submitted, out of extracted affiliations of authors, on average 72.99% were actually correct regardless the PDF, hence treating all PDFs equally. Micro-average recall means that 64.56% of the time when a PDF is submitted, the affiliations of authors were correctly identified. Micro-average recall is calculated by aggregating the contributions of PDFs and by not treating all PDFs equally. Macro-average recall means that 62.59% of the time when a PDF is submitted, the affiliations of authors were correctly identified regardless the PDF, hence treating all PDFs equally.

Upon evaluation, I noticed that sometimes the output of the GROBID has duplicate affiliations. It happens when an author is associated with two different departments in the same institution. The development team suggested to remove duplicate

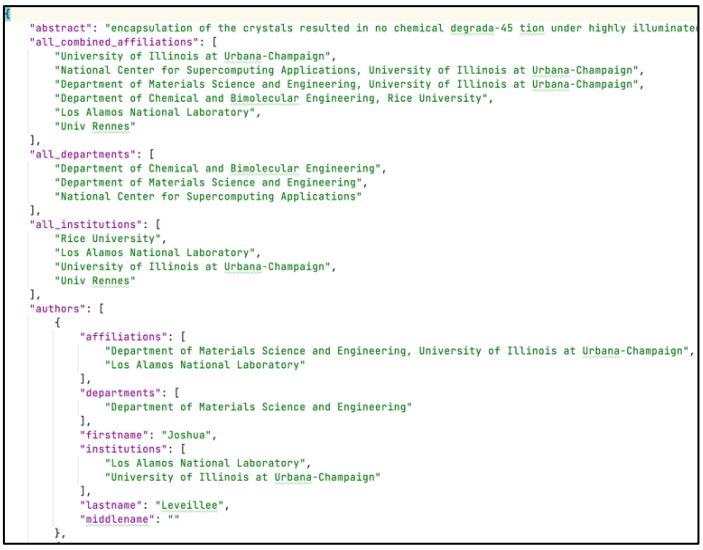

Fig. 1 JSON Response of Feature Extraction

affiliations. Sometimes, GROBID identifies institutions as departments. Since I was only extracting institutions, when an institution is misclassified as a department, the output does not contain it. Upon discussing with the team, we decided to include both department and institution in the affiliations. Then, I modified the JSON output to include all combined affiliations list in the form of department, institution, all departments list, and all institutions list. I also modified the authors object to include three lists of affiliations in the above mentioned format (all combined affiliations, all departments, and all institutions).

III: Progression

Prior to the internship, my knowledge on open-source, machine-learning based tools such as GROBID and SciWING was limited. In addition, I was not aware about the review and release systems such as RASSTI. During the internship program, I learned about machine-learning based tools for parsing headers of scientific documents. I also implemented various solutions using GROBID, SciWING, and PyPDF2 and performed multiple evaluations on speed and accuracy. I also learned about RASSTI and how it populates library systems with bibliographic data. Therefore, I understood the problem and its challenges. It helped me design and develop a better performing solution.

Moving forwards, I created two web services using the GROBID extractor: */extractfeatures* and */extractfulltext*. */extractfeatures* web service is capable of uploading a PDF document to extract features (titles, authors, keywords, and abstracts) and send the JSON response with the extracted features. */extractfulltext* web

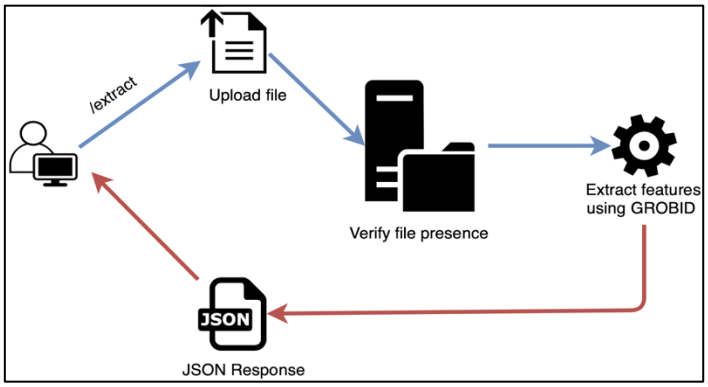

Fig. 2 Extract Service Overview

service is capable of uploading a PDF document to extract titles and full text including abstracts and send the JSON response with the extracted full text. I created these web services using the Flask framework. I used postman to invoke */extractfeatures* and */extractfulltext* web services by sending research papers as files and displaying the JSON response from them.

IV: Academic Relevance

Course work

During this internship program, I initially explored available machine-learning based tools for feature extraction on PDF documents and evaluated them to find a tool that works the best. For the evaluation, I curated a dataset with 20 sample PDFs and their features. Based on the evaluation, we decided not to utilize PyPDF2 (did not extract the complete list of authors or abstract or keywords) and SciWING (extraction was not real-time) and proceeded with GROBID to extract metadata from PDF documents in real-time as it was fast and accurate compared to PyPDF2 and SciWING. I also looked into multiple ways of structuring the extracted data. I initially created a tabular format to store the data. Since it did not provide the best structure to hold nested information, I proposed to use a JSON object to structure the extracted data. After discussing this issue with Brian and the development team, we decided to use a JSON objects to structure the output of GROBID. After multiple rounds of discussions with the development team, we agreed on which fields should be present in JSON objects. As the next step, I implemented two web services functions to upload a PDF to extract features and full text from the uploaded PDF, using the Flask framework. Upon extraction of features and full text, these web services return JSON responses. I used postman to invoke these web services and to display the JSON responses from them.

From completing this project, I learned about multiple key aspects of real world applications. They are speed, accuracy, and completeness. To fulfill the requirements of the project, I evaluated multiple tools and decided tools/technologies to proceed with based on the results. This practical experience contributed to my understanding of importance of speed, accuracy, and completeness when comes to real-world applications.

Professional Literature

PDF has become the most commonly used mechanism to share research publications. The metadata such as title of paper, authors, affiliations of authors, abstract, and keywords are usually included in the header of the paper. One study compiled a table of existing header parser tools [1] and examined their top-level behaviors. The authors found GROBID, ParsCit, Header Parser Service and Mendeley are header parser tools that use machine learning algorithms. However, the study [1] evaluated the performance of parser tools using ten real hand-picked (not entirely random) research papers by conducting an informal test and a side-by-side comparison of the results. The authors of the study [1] did not present an objective evaluation on machine learning based header parser tools. Therefore, I evaluated two machine learning based tools the study [1] identified (GROBID and ParsCit) in terms of accuracy and speed. We used SciWING as ParsCit is integrated in SciWING. Between GROBID and SciWING, SciWING did not pass the speed requirement of our project. Therefore, we did not utilize SciWING.

Another study [2] proposed a hybrid method for the extraction of header information from the research papers using GROBID, ParsCit and Mendeley, since no single machine-learning based tool gives 100% results against all sample research papers when parsing the header. The authors of the study [2] have merged the results of the three tools to achieve accurate header extraction.

They have applied the proposed method on 75 sample research papers and achieved an overall accuracy of 95.97%.

Our project requirement was to aid users by populating information in forms when entering information into the LANL's review system. Therefore, the web service that I developed during this internship program is capable of extracting metadata such as title, authors, affiliations of authors, keywords, and abstracts from PDFs submitted via LANL's review system, and populating information in forms using the JSON response of the web service. Hence, users will not have to manually enter metadata when submitting PDFs.

Since the real-time metadata extraction from submitted PDFs could be utilized in more scenarios apart from LANL Research Library systems, I plan to explore this idea in my future research such as populating configuration files from the extracted metadata from PDFs related to sensory devices. In addition to the research and development, the practical experience I gained throughout this internship program is crucial for the improvement of my ongoing research work and future career development.

References

[1] Kevin Yao, Mario Lipinski, Bela Gipp and Jim Pitman. "Header Extraction from Scientific Documents".

[2] Saleem, Ozair, and Seemab Latif. "Information extraction from research papers by data integration and data validation from multiple header extraction sources." Proceedings of the World Congress on Engineering and Computer Science. Vol. 1. 2012.

V: Future Projections

This is my first internship program within the USA, outside graduate assistantships within ODU.

Research and Development Engineer, Software Developer, Back-End Web Developer, and Data Scientist are some of the career options available in the field of this internship program. During this internship, I learned how to extract metadata from PDFs using machine-learning-based tools, and how to structure the extracted features. I objectively evaluated and chose a machine-learning-based tools and chose the most appropriate tool based on the results. To succeed in research and development related career, we need to explore and evaluate multiple tools/techniques before making a decision. Practical experience gained through research and development is necessary to obtain a career in the future.

Also, this internship program provided me the opportunity to network with professionals in LANL working on projects related to Computer Science. These connections will be beneficial when I will be actively looking for job opportunities in the future. I also identified the importance of working as a team in order to successfully and efficiently contribute to a project. This internship program adds a valuable experience entry to my resume which will help me find one of the aforementioned career opportunities once I graduate. I have updated my Resume, and CV to reflect the experience I gained from this internship program.

VI: Conclusion

The overall experience I gained by working as a Summer Research Intern at LANL within 12 weeks, is an incredible work experience. Even though I worked remotely due to the social distancing restrictions because of Covid-19 pandemic, the entire ISC team was supportive. I would recommend both graduate and undergraduate students to apply for internship positions since the experience I gained through this internship is crucial for my future career.

I would like to express my gratitude to my PhD advisor, Dr. Sampath Jayarathna, for encouraging me to apply for Summer internship programs, to my internship supervisor, Brian

Cain, for guiding me throughout the internship by providing feedback and suggestions, and to Dr. Martin Klein, for recommending me for this position. I am thankful for the opportunity to work at LANL as a Summer Research Intern with Brian Cain and the ISC team.

VII: Beneficial Suggestions

We did not submit a proposal based on the solution to any problem. However, during my internship, I proposed a few solutions based on my research and development. For instance, after exploring and objectively evaluating some of the available machine-learning based tools (PyPDF2, SciWING, GROBID), we decided to proceed with utilizing GROBID to extract metadata from PDF documents in real-time based on speed and accuracy. For structuring the extracted data, I created a tabular format to store the data and also proposed to use a JSON objects. After discussing with the development team, we decided to use a JSON objects to structure the output of GROBID. I also proposed to have the feature extraction as a web service to make it easy to integrate with the RASSTI system.

VIII: Photos/Quotes/Video

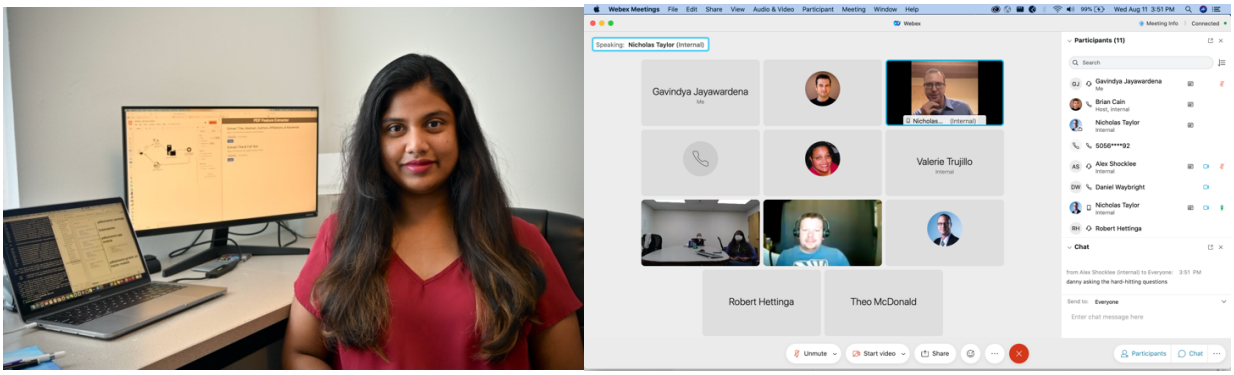

Fig. 3 Photos of me working remotely as a Summer Intern at LANL